Metaphors, Mantras, and Mnemonics: Communication Competencies in Educational Measurement

Metaphors, Mantras, and Mnemonics:

Communication Competencies in Educational Measurement

Andrew Dean Ho, *Harvard Graduate School of Education*

Forthcoming in *Educational Measurement: Issues and Practice*

April, 2026

Abstract

I provide evidence and arguments for increased investment in communication competencies in educational measurement. Building on the 2024 framework for Foundational Competencies in Educational Measurement, I emphasize communication as both a distinct domain and an interactive amplifier of collaboration, technical, and measurement competencies. I draw empirical and conceptual support from labor economics, psychometric job postings, and longstanding calls within measurement scholarship.

I then propose a conceptual toolkit called the "3-4-5 Triangle" of measurement mnemonics: the three Ws of test context, the four quadrants of testing purpose, and the five Cs of validity evidence. I illustrate this toolkit in teaching, research, and policy engagement, and situate it within broader commitments to collaboration, community engagement, and transformative research. I conclude by arguing that systematic cultivation and assessment of communication competencies are necessary for the continued relevance, impact, and legitimacy of educational measurement.

Introduction

Since 2022, my colleagues and I have been engaging our field in conversations about Foundational Competencies in Educational Measurement (Ackerman et al., 2024; Ho et al., 2024). Our 12-member Task Force, charged by then-president of the National Council on Measurement in Education (NCME) Derek Briggs, identified three competency domains: 1) communication and collaboration competencies, 2) technical, statistical, and computational competencies, and 3) educational measurement competencies. Within this third domain, we then identified five overlapping and intersecting subdomains (Ackerman et al., 2024, Figure 2). In this article, my focus is on the first domain of Communication and Collaboration, and on Communication, in particular, including its intersections with other domains and subdomains.

My fellow Task Force members and I have been presenting this framework to members of sibling organizations, including the Council of Chief State School Officers, the Psychometric Society, and the International Test Commission. This journal also published eight commentaries on our article that constructively critiqued various elements of our framework. We summarized and discussed these in a rejoinder (Ho et al., 2024). Through all of these interactions, I have found it notable that commenters only critiqued our emphasis on communication competencies for being insufficiently specific and participatory, a call for more, not less (e.g., Brookhart, 2024; Crespo Cruz et al., 2024; Middlebrook et al., 2024; van Orman et al., 2024).

In one of these commentaries, Ying Cheng and her colleagues coded 80 job postings related to psychometrics. They found empirical support for our framework in what

employers state that they seek. In a compelling graphic, they found that communication and collaboration competencies, what Cheng and her coauthors described as "soft skills," accounted for almost as many of the keywords in job postings as measurement itself, and five times more than the current hot topic, Artificial Intelligence (Cheng et al., 2024, Figure 8).

Their findings mirror findings in labor economics. Deming (2017) shows that jobs requiring high levels of social interaction grew by nearly 12 percentile points as a share of the U.S. labor force. Deming (2022) also argues that how we develop social skills is a largely unexplored frontier, with only a few promising randomized experimental interventions. For example, an intervention in the garment industry in India that focused on communication and other social skills increased productivity by 13.5 percent (Adhvaryu et al., 2023). Communication is a skill we can improve in ways that improve our work.

Braun and Mislevy (2005) have also long argued for communication competencies. They identify nine intuitive principles, or "p-prims," that they say are good enough to guide everyday action, but, to follow a physics metaphor, not send someone to the moon. They close their article saying, "we need to do a much better job of communicating to a variety of audiences the basics of testing and the dangers we court when we ignore the principles and methods of educational measurement. Communication is a form of teaching, and we should take the challenge of this kind of teaching more seriously than ever before" (p. 497).

With such consensus, a central question is how to support communication competencies in our field. The remainder of this article supports three claims:

1) The most important communication competency in educational measurement is: all the other competencies. We cannot communicate effectively about educational measurement if we do not first understand educational measurement. However, building communication is not zero-sum. Time spent building communication competencies should not detract from building educational measurement competencies; it should reinforce them. I grant the old saying, “if you can’t explain it simply, you don’t understand it yet.” However, I would add, “trying to explain it, helps you understand it.”
2) We can improve communication within our field and beyond by working towards shared metaphors, mantras, and mnemonics. I present examples of these below.
3) Finally, this emphasis on communication enables us to a) communicate precisely about imprecision, a longstanding strength of our field, and b) communicate decisively amidst indecision, a competency that I believe we in our field need to develop further.

Positionality

My arguments about communication competencies in measurement are shaped by my personal and professional trajectories. I establish my positionality here because positionality helps to contextualize the questions we ask, the findings we highlight, and the conclusions we draw. This is true not only in qualitative research but also in quantitative and measurement research (Castillo & Gillborn, 2025; Gillborn et al., 2018; Milner, 2007)

and therefore in this paper. Core to my positionality is my upbringing in Hawai‘i and Japan, and my experience being raised by and then becoming a teacher.

I am a fifth-generation Chinese American who grew up in Hawai‘i in comparatively privileged circumstances, largely protected from the kinds of minoritization that many students of color experience in U.S. schools. I have both benefitted from and been constrained by the “model minority” myth (Lee, 2009), which positions Asian Americans as high-achieving and well-behaved but lacking agency and power. My own experience of having test scores interpreted as evidence of promise likely shapes my commitment to creating similar opportunities for others.

As an undergraduate studying abroad in Japan in 1998, I interviewed high school students and teachers about their educational opportunities. Their often negative accounts of testing shaped my belief that high-stakes uses of tests can constrain curriculum and pedagogy, deprofessionalize teachers, dehumanize students, and increase cynicism and fatalism. These experiences inform my current stance as a psychometric pragmatist: I understand test scores as imperfect and frequently misused, but I also believe they are tools worth improving and contextualizing rather than abandoning.

My mother and grandmother were teachers. I began teaching formally in my first summers in high school, and I have taught formally almost every year since then. Because I am a teacher, I came to testing and measurement indirectly, in part because I have witnessed firsthand the power of testing and quantification to focus and inspire learning, as well as to narrow and discourage learning. As I observed in a recent commentary reviewing the work of 12 NCME Past Presidents and their coauthors (Ho, 2024), I believe

many come to this field because we recognize and experience both the power and peril of test scores. This motivates my focus on communication competencies as a lever for improving uses and interpretations of scores, as well as preventing predictable unintended consequences of common test uses.

The Testing Expert's Monkey's Paw: "May our test scores be used"

One of my mantras for my statistics and measurement students is, as I noted above: "communicate precisely about imprecision." I believe we in our field are generally skilled at this. Because of our statistical and psychometric training, and because others tend to be, as I describe, "weak to numbers" (see also Porter, 1996), it might bring us some satisfaction to pull back the curtain, like a mythbuster or a magician revealing a trick. "Look," we might say, "here was the bias and the standard error of these numbers, lurking here the whole time!" In this way, we are trained to be and might enjoy being "precise about imprecision." But there is a second mantra where I think we fall short: "communicate decisively amidst indecision." I know we as measurement experts can do the first. My encouragement to those in our field, to communicate better, is to do more of the second.

In an old commentary, I illustrated this imbalance using a metaphor that test scores are like Lego blocks. I wrote:

> I imagine our community of modern testing experts as Ph.D.-level designers of the snap- together children's blocks known as Legos. We lovingly construct versatile pieces meant to cohere elegantly into castles and starships, with multistep instruction manuals detailing their appropriate use. And we sit back and shake our

> heads as Lego users neglect our instructions and build, instead, nondescript multicolor towers or, if we're lucky, an occasional car. We think we are selling castles and starships, but our product sells less for its intended purposes than its flexibility; its perceived utility exists in the eyes of beholders. (Ho, 2016, p. 34)

As the metaphor suggests, I think we become protective of test scores and circumscribe their use to elaborate and specific contexts, given what we know about score imprecision and historical and ongoing score misuse. But communication requires us to meet users where they are. We must do more than provide elaborate instruction manuals that no one reads or uses.

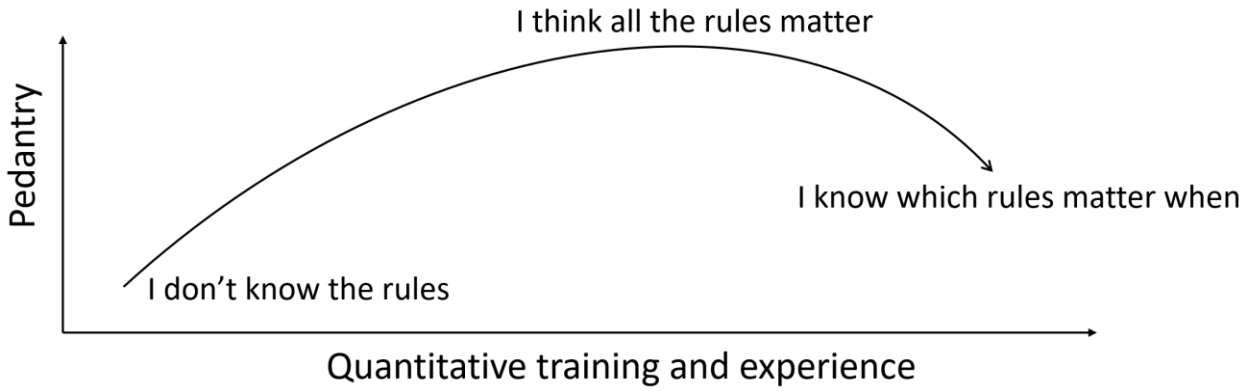
In the story of the Monkey's Paw (Jacobs, 1902/1906), a charm grants wishes with horrible unintended consequences. It leaves us with the impression that a statement like, "may your wishes come true," is not a blessing but ominous, and ultimately a curse. Similarly, for testing experts, it sometimes feels like the worst curse imaginable is to say, "may your test scores be used." However, I believe it is part of our job to communicate when and how, given imprecision, we and others *should* interpret and use test scores.

The Pedantry Parabola

Because measurement experts have accumulated experience with test score misuse, our guidance tends to err on the side of caution when we teach and provide technical advice. As I have argued, I think we are often too cautious, and this over-caution can impede practical work. To help students navigate this tension, I introduce a concept called the Pedantry Parabola (Figure 1), where pedantry refers to overemphasis on minutiae

regardless of their practical consequences. The parabola describes a pattern in how expertise develops: we begin without knowing the rules; we learn rules and believe them to be inviolable; then we learn which rules matter when, and why.

Figure 1. *The pedantry parabola*.



For example, in my statistics class, I have a lecture called, "the regression assumptions that matter." The title refers to the fact that some regression assumptions that I myself spent weeks learning when I was a student do not matter for most practical purposes in social science research. For example, the assumptions of homoskedastic and normal error distributions are both generally addressed by heteroskedasticity-robust standard errors (Hayes & Cai, 2007; Huber, 1967; White, 1980) and the Central Limit Theorem (e.g., Gelman & Hill, 2006), respectively. Students who have climbed the parabola may recognize this stage in themselves: obsessive and ultimately inconsequential perseveration over scatterplots and histograms. And if astute readers feel quick to caution against this permissiveness, for example, when sample sizes are small (MacKinnon et al.,

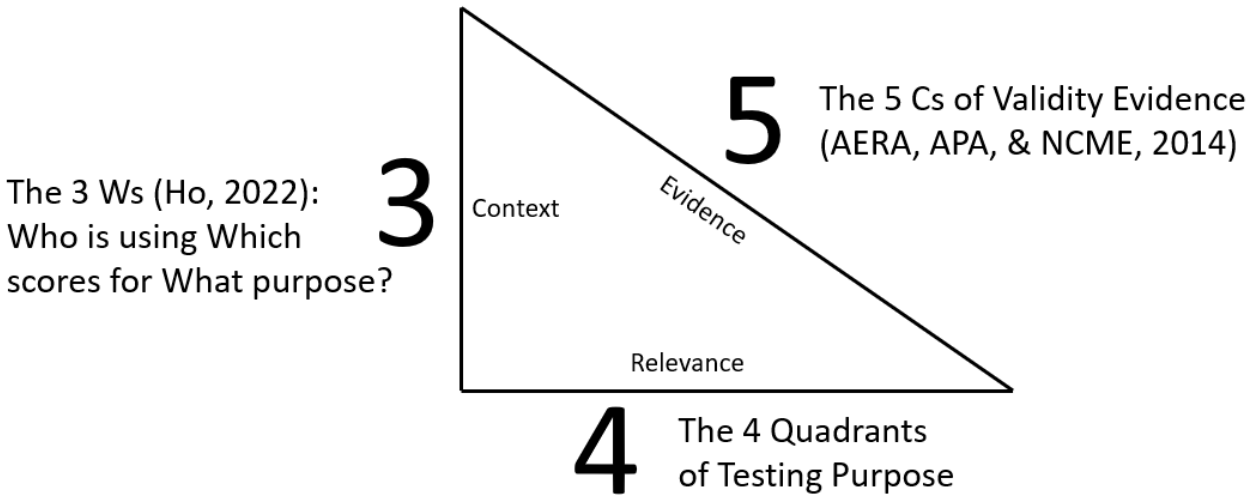
2023; Wilcox, 2012), that reaction is itself an illustration of navigating back down the parabola toward practical judgment in context.

I joke with my measurement students that the answer to every question on my assignments is, "it depends." Then I stop joking, and I say: "your job is to tell me what it depends on, and then give me your final recommendation." Good communication, as Braun and Mislevy (2005) observed, is a form of teaching. Good teaching ultimately requires both explaining why the answer depends on context and advocating clearly for an evidence-based course of action. Navigating the pedantry parabola ultimately means knowing which mode a given moment demands: when to slow down and build understanding, and when to make the case and accept the costs of simplification. Both are communication competencies worth cultivating: to be precise about imprecision, and to be decisive amidst indecision.

The 3-4-5 Triangle of Measurement Mnemonics

In order to get over the pedantry parabola, to learn and teach which rules matter when, and to improve our communication with those inside and outside the field, I present here my "3-4-5 triangle of measurement mnemonics," with thanks to Pythagoras (Figure 2). This is a mnemonic for other mnemonics, a meta-mnemonic. Each mnemonic does not represent new ideas as much as re-expresses existing principles in, I hope, memorable ways. They have helped me to teach, and I have found that they help me and my students, and my colleagues who know these mnemonics, communicate better among ourselves. Together, I hope these mnemonics help to make measurement memorable.

Figure 2. *The 3-4-5 triangle of measurement mnemonics.*



The 3-4-5 triangle comprises the 3 Ws: Who uses Which scores for What purpose (described first in Ho, 2022), the 4 quadrants that identify common testing purposes (described first in Ho, 2024), and the 5 Cs of validity evidence, a mnemonic for the five sources of validity evidence in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). The 3 Ws help us to understand the context of measurement. The 4 quadrants help us to understand which evidence is most relevant in which contexts. And the 5 Cs help us to remember the sources of evidence that matter. Together, these three mnemonics move from context to relevance to evidence (Figure 2). I address each in turn.

The 3 Ws: Who is using Which scores for What purpose

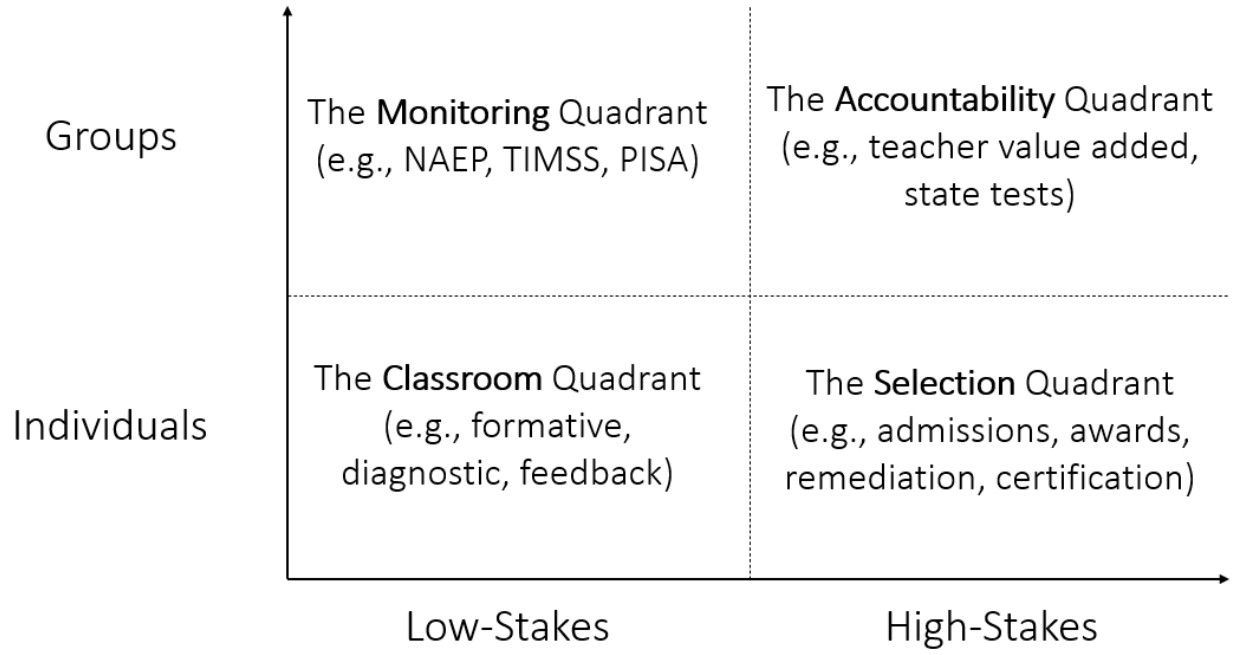
It is not possible to develop any research or validation agenda in testing without first asking and answering "the 3 Ws question." Standard 1.1 is our first standard for a reason, "The test

developer should set forth clearly how test scores are intended to be interpreted and consequently used" (AERA, APA, & NCME, 2014, p. 23). I prefer the 3 Ws to Standard 1.1, and not simply because the alliteration may be memorable. The *Standards* use passive voice: "how test scores are intended to be interpreted and consequently used" (p. 23). This raises important questions: Interpreted by whom? Used by whom? I have encouraged the Joint Standards Committee that is currently revising our standards to use active voice whenever possible. Who are the actors who have the power to use test scores to make decisions? They must be the ones who take responsibility for weighing validity evidence. The 3 Ws appropriately focus us on these actors, the specific scores they use, and the context for evaluating which evidence matters for their intended score use.

The Four Quadrants of Testing Purpose

To encourage users to consider how required test score evidence depends on purpose, I use another mnemonic: the Four Quadrants. This simple framework distinguishes between high-stakes and low-stakes uses of tests, on one dimension, and individual and aggregate uses on the other (Figure 3). Each quadrant represents a common answer to the 3 Ws question, defined by who typically uses scores, at what level of aggregation, and with what stakes.

Figure 3. *The four quadrants of testing purpose.*



Note. NAEP = National Assessment of Educational Progress; TIMSS = Trends in International Mathematics and Science Study; PISA = Programme for International Student Assessment.

In the classroom quadrant, teachers, parents, and students use scores for low-stakes formative purposes. In the selection quadrant, admissions officers and school personnel use scores for higher-stakes admission, selection, and remediation. In the accountability quadrant, policymakers and agency officers use aggregate scores for teacher, school, or district accountability. In the monitoring quadrant, policymakers and the public use lower-stakes aggregate scores for periodic assessment of educational standing and progress.

The Four Quadrants aid communication by reminding us to establish context before evaluating evidence. Must student test scores have a reliability of 0.9 if the goals are

classroom feedback or aggregate monitoring? Shouldn't guardrails against cheating differ depending on whether stakes are at the school level or the student level? The quadrants also help the public put testing debates in perspective: the vast majority of testing time is spent in the classroom quadrant, while the vast majority of testing controversy centers on the accountability and selection quadrants.

The Four Quadrants are deliberately simple and lead to three natural extensions. First, the axes are continuous. There is no clean line between low stakes and high stakes, and actors aggregate scores to different levels. Locating tests in this continuous space, rather than assigning them to a single quadrant, motivates more precise answers to the 3 Ws question. Second, test users cross quadrants. Actors routinely use the same scores for different purposes. Third, this *purpose drift* is predictable. Many who design tests for classroom purposes find their scores repurposed for selection or accountability. We must collect validity evidence to anticipate this movement rather than follow it (Ho, 2014; Ho & Polikoff, 2025).

This literature establishes principles more nuanced than any mnemonic can represent fully. The Four Quadrants strike a useful balance between accuracy and interpretability. A good mnemonic should not conflict with more complex and accurate frameworks; it should inspire them.

The 5 Cs of Validity Evidence

Once we have established context through the 3 Ws and identified the corresponding quadrant, the validity evidence we produce and evaluate should follow the five sources outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME,

2014). Table 1 presents a mnemonic re-expression of these sources as the "5 Cs of Validity Evidence." The alliteration is forced but functional: Content (or Construct), Cognition, Consistency, Correlation, and Consequence map directly onto the five sources. The goal is to be memorable, not novel.

Table 1. *The 5 Cs of validity evidence*.

The 5 Cs	AERA/APA/NCME (2014)
Content / Construct	Evidence based on test content
Cognition	Evidence based on response processes
Consistency	Evidence based on internal structure
Correlation	Evidence based on relations to other variables
Consequence	Evidence based on consequences of testing

Note. The 5 Cs are a mnemonic re-expression of the five sources of validity evidence in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Nonetheless, the 5 Cs can illuminate priorities that the Standards' own presentation can obscure. The idea that Content (or Construct) is a foremost guide to all validity judgments can be reduced to a mantra: First C, first. The ordering of the Five Cs roughly reflects the sequence of steps in test development and validation. And presenting the Five Cs together holistically can reveal what researchers over- and underemphasize in validation practice.

For example, as they learn the Five Cs, students remember that no amount of internal structure evidence can compensate for content alignment. They recognize the chronic underuse of cognitive response processes in validation practice. They see how the

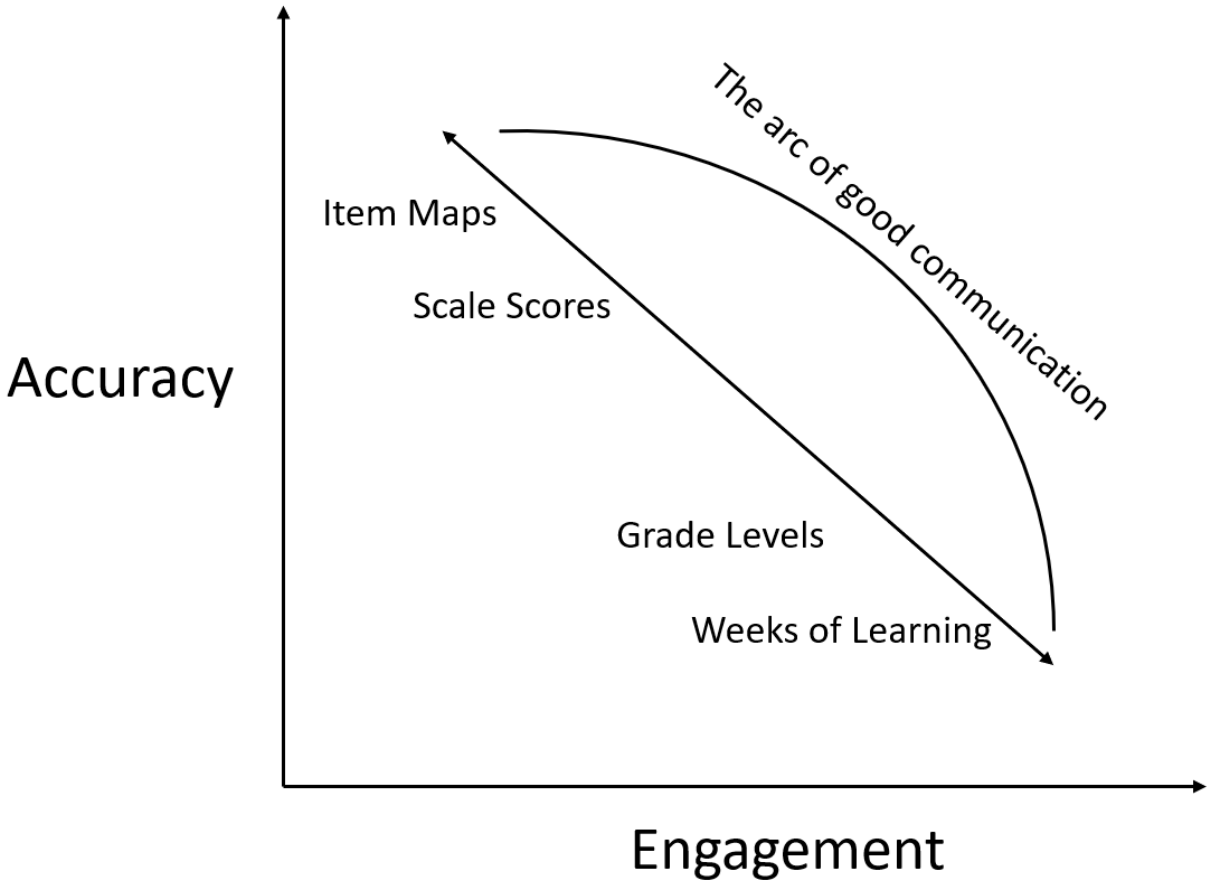
perceived importance of correlation, the fourth C, is inflated by misleading terms like "concurrent validity" or "predictive validity." Consequence, the fifth C, asks what the causal effects of testing (or not testing, or testing differently) actually are. They see that this evidence requires applying methods for causal inference that our field too often leaves to economists and policy researchers. I believe we should be engaging more with these questions ourselves.

My students have mistakenly cited these as "Ho's 5 Cs," as if I had authored the 2014 Standards or could claim more than a century of validity theory. This shows the mnemonic did its job. (For the record, please cite these as AERA/APA/NCME, 2014.) Good mnemonics deepen learning. And when we share mnemonics, they create a shorthand that improves communication and expands the community of experts who engage productively with validity evidence.

The Accuracy-Engagement Tradeoff

Measurement experts face a persistent tension in communicating results to various audiences. Figure 4 illustrates this as an accuracy-engagement tradeoff. Technical reporting formats like scale scores and item maps preserve accuracy but are inaccessible to most audiences. More accessible formats like grade levels and "months of learning" engage broad audiences but invite misinterpretation. The diagonal line illustrates the tradeoff. Accuracy and engagement pull us in opposite directions. Moving toward one often means sacrificing the other.

Figure 4. *The accuracy-engagement tradeoff in test score reporting.*



The 3-4-5 triangle does not dissolve this tension, but it helps us navigate it. By first establishing who is using which scores for what purpose, and by identifying which quadrant a given use occupies, we can make defensible choices about where on this diagonal to engage audiences. A researcher using aggregate scores for monitoring purposes in the monitoring quadrant should evaluate evidence differently than the developer of an admissions test whose scores may discourage an applicant from applying to a school or an admissions officer from selecting a student. Knowing the context does not eliminate the tradeoff. It turns an arbitrary choice into a principled one.

The deeper argument of this paper is that good communication competencies can bend this diagonal. We can call it an arc of good communication, moving us toward reporting that is both more accurate and more engaging. We can be precise about imprecision by communicating the limitations of any reporting format honestly. We can be decisive amidst indecision by choosing a format based on evidence about context and audience, and by framing that choice to anticipate and discourage predictable misuse. The two examples that follow, the development of the Stanford Education Data Archive (SEDA) and the choice of grade-level reporting, illustrate how the 3-4-5 triangle guided both reporting decisions and their communication to colleagues and the public.

The Stanford Education Data Archive (SEDA)

For over 15 years, my colleagues Sean Reardon, Erin Fahle, Ben Shear, and many others have collaborated to develop SEDA, a national database of linked aggregate-level test scores comparable across states and over time (Reardon et al., 2024). The effort appears to challenge a decades-long consensus that we should not use National Assessment of Educational Progress (NAEP) to link state test scores (Feuer et al., 1999).

The 3-4-5 triangle provided the framework for justifying and communicating this decision. The 3 Ws establish the context: researchers and the public are the who, district average scores are the which, and monitoring patterns of achievement and inequality is the what. The monitoring quadrant (Figure 3) is the relevant context for evaluating evidence. We provided validity evidence aligned with the third C, consistency: a study demonstrating how well an aggregate linking method recovers district NAEP scores for large urban districts (Reardon et al., 2021). We were precise about imprecision. Then, we were decisive amidst

indecision. Since 2015, we have been releasing aggregate scores regularly and publicly at edopportunity.org.

In a special issue of the *Journal of Educational and Behavioral Statistics*, colleagues commented on our methods and decisions with varying degrees of concern. Moses and Dorans (2021) offered a pointed critique, citing my own prior work: “In absence of further studies, we recommend that the validation efforts of the Reardon et al. linking procedure be treated with scientific skepticism and accompanied with the warnings that have been expressed about similar studies where state assessments are linked to NAEP (e.g., Ho & Haertel, 2007)” (p. 200). The citation of my earlier cautions against my later decisions is a fair challenge that deserved a direct response.

The 3-4-5 triangle provides one. In our rejoinder, we acknowledged that “we largely agree with our colleagues about the evidence and what the evidence means” (Ho et al., 2021, p. 210). We nonetheless defended our conclusion "because we believe our colleagues are underestimating the benefits and overestimating the potential for downstream harm… Possible misuse is not misuse" (p. 211). The 3-4-5 triangle does not resolve such disagreements but clarifies them. It shows how colleagues can acknowledge the same evidence and still reach different conclusions. Context, purpose, and evaluating benefits against risks are often judgments that evidence alone cannot determine.

Grade-Level Reporting

Grade-level reporting has a long history in educational measurement (Hoover, 1984), and a long history of controversy. Grade-level scales fell out of favor for student-level reporting because many misinterpreted results as prescriptions for grade retention or promotion.

These concerns apply less straightforwardly to aggregate monitoring, where there are no direct inferences about any individual student. NAEP, for example, reports only aggregate scores, and the relevant question is not whether a student is "on grade level" but whether aggregate trends are moving in the right direction and by how much (Ho, 2023). I would nonetheless caution against grade-level reporting of achievement gaps between sociodemographic groups, as framing disparities in grade-level terms can make them seem more immutable and reduce investment in equalizing educational opportunity (Quinn & Desruisseaux, 2022).

My colleagues and I use grade-level reporting in SEDA. Using a calculation we describe in our technical documentation (Fahle et al., 2024), we estimate that district average test scores span roughly six grade levels across the United States, a substantial range. We could express this as two standard deviation units, and indeed the correlation between key results from the two scales exceeds 0.99. One is essentially a linear transformation of the other. We opt for grade levels because the metric is more accessible to researchers and the public. The norm underlying this scale is explicit: one grade level equals one-quarter of the distance between national average Grade 4 and Grade 8 NAEP scores for specified cohorts in the 2010s (Ho, 2023). Compare this to Cohen-type effect sizes in standard deviation (SD) units. These are also norm-referenced. The norm is the SD of whatever reference population the researcher selects. SD units are not inherently more objective than grade levels. They are more familiar to researchers and less familiar to everyone else.

Betebenner and DePascale (2024) raise legitimate concerns about time-based reporting, critiquing the semantics, technical foundations, and content-referencing of time-based metrics. I concede many of their concerns. Their semantic and content critiques apply to most reporting formats, not just time-based metrics. Their technical critiques carry more weight at the individual student level than at the aggregate level where SEDA and NAEP operate. The 3-4-5 triangle clarifies our disagreement. In the monitoring quadrant (Figure 3), the benefits of engagement are substantial, and good communication competencies can encourage appropriate use. When the public response to historic pandemic-era score declines is, "is 3 scale score points a big deal," a linear transformation with a clear norm group is a defensible communication choice (Ho, 2023).

Communication Competencies are Learned

Learning to communicate well takes years, and I hope early errors are more instructive than discouraging. As one illustration, early in my career, I described my research on growth models to a radio reporter in Iowa. In a live interview, I explained that nonparametric statistics provided a useful foundation for understanding how nonlinear monotonic transformations could affect growth estimates. The reporter wrapped up quickly and moved on to sports. I learned a lesson: know my audience, meet them where they are, and set realistic communication goals. I later developed metaphors: Comparing growth with test scores is like comparing the height of two children on pogo sticks. This was imperfect but hopefully intriguing enough to pull a listener toward understanding.

A more consequential lesson came later, on the National Assessment Governing Board (NAGB), which governs NAEP. As a new board member in 2012, I had spent the prior five years arguing against proficiency-based reporting of trends and achievement gaps (Ho, 2008). Sitting at a table with governors, legislators, teachers, and business representatives, I remember thinking it would be appropriate to make that technical argument at length. It did not land well. My fellow board members heard not a nuanced argument about the limitations of proficiency percentages but a challenge to proficiency standards that are a cornerstone of NAEP. I had climbed the pedantry parabola (Figure 1) and failed to navigate back down.

Over years as a board member, I learned to translate technical arguments into decisive recommendations without losing their precision. When the NAEP transition from paper to digitally based assessments in 2017 threatened to disrupt the trend line, I argued that NAEP had one job: to measure educational progress. The question was not whether to maintain the trend but how. The precise answer required a robust bridge study, which the National Center for Education Statistics (NCES) led and reported. This evidence supported the NAGB/NCES decision to maintain the trend across the transition. We needed to be precise about imprecision and then decisive amidst indecision. When my NAGB colleagues gifted me a cape that said “protector of trend” at the end of my two terms, I accepted it in good humor as a sign that my communication competencies had improved.

Assessing and Rewarding Communication Competencies

If we value communication competencies, we must create conditions that encourage students and early-career scholars to develop them. Every semester, I administer what I call the Conversational Celebration of Learning (CCL), a 20-minute oral examination, to every student in my measurement and statistics courses (Ho, 2025). Each CCL draws from assignments students have already completed, so there are no surprises at the outset. I ask students, for example, to define reliability, first as a concept they would explain to someone on the street, then more precisely as a correlation, then as a percentage of variance. Could you do this right now, without a text or a search engine? Without AI?

This is harder than many expect, because students often have little practice speaking about measurement. The CCL incentivizes this practice. Students prepare by having conversations with each other, not just by reviewing notes alone. As large language models can make written assignments less trustworthy, oral exams also offer a direct measure of student fluency that AI cannot easily substitute. I encourage teachers of measurement to adopt similar assessments. Let us measure what we value.

I also argue that tenure criteria should incentivize communication competencies. My institution requires evidence of impact along four criteria, i) research and scholarship, ii) teaching, iii) relevance to education policy and practice, and iv) institutional contribution. Communication competencies support all of these. Clear communication to policymakers and practitioners, with evidence of uptake, should support a case for tenure. This is especially true for our field: NCME is the National Council on Measurement *in Education*, and our work is most fully realized when it informs educational policy and practice.

This requires institutional will, and it often requires individual courage. In 2011, I asked my mentor Ed Haertel whether I should accept a colleague's nomination to NAGB as a sixth-year assistant professor. He replied thoughtfully: "I hope that you serve on NAGB at some point in your career, but I would not recommend this kind of service for anyone at your career stage" (E. Haertel, personal communication, 2011). Five weeks later, he wrote again: "I advised you against serving on NAGB, but your name keeps coming up in various conversations… I'm still ambivalent… but I'm more inclined than I was to believe that you might really be able to make a positive difference if you were willing to serve" (E. Haertel, personal communication, 2011). To early-career scholars: if a communication opportunity arises that positions you to make a meaningful difference, I hope you take it, even if your competencies are still developing. To more established members of the field: I hope we reward that effort and its impact. Good communication is good for our field.

Beyond Communication: Limitations and Future Directions

The most important limitation to emphasize is that communication, however improved, is necessary but not sufficient. Domain 1 of the Foundational Competencies framework is not communication alone but communication *and collaboration* (Ackerman et al., 2024, emphasis added). This article focuses on speaking and explaining, that is, on transmitting expert knowledge more clearly. But competencies that matter for transformative research also include listening, learning from communities, and co-designing research with the people it is meant to serve.

In the Spencer Foundation Task Force on Preparation for Transformative Research (2025), my colleagues and I argued that graduate preparation in education must develop capacities for community engagement, theoretical and methodological pluralism, ethics, digital technologies, and knowledge mobilization. The implication is that measurement research cannot positively transform school communities absent genuine collaboration. Scholars studying culturally grounded assessment practice illustrate what this collaboration can look like in practice. Shultz and Englert (2021), for example, describe a community-engaged validity framework developed with Native Hawaiian language immersion students. No mnemonic can substitute for this kind of engagement.

Three additional limitations deserve brief acknowledgment. First, the framework I develop here arises from my own experience as a practitioner and scholar focusing primarily on U.S. educational measurement. Although I have found these tools successful in international presentations, different policies and contexts may require adaptation of, for example, the four quadrants, or the mantra to be decisive amidst indecision. Second, as befits many presidential addresses, my arguments draw heavily from autobiographical anecdote. This risks survivorship bias: I chose to illustrate the failures I survived. This is not strong evidence that others should fail similarly or that the risks I took are appropriate for everyone. Third, as I stated in the first claim in this article, good communication cannot substitute for measurement expertise. If mnemonics circulate without the underlying knowledge they encode, they risk encouraging confident misuse. The 3-4-5 triangle scaffolds expertise. It cannot replace it.

Conclusion: Making Measurement Memorable

One goal of expert communication is changing minds in ways that last. This is also the goal of good teaching. In 2013, I inherited a beloved statistics course from John Willett, an educational statistician who had accumulated what I then considered a vast collection of teaching gimmicks. One of them was what he called potato technology: giant dowels stuck into a potato to illustrate how Principal Components Analysis summarizes correlated variables. At the close of the semester, he would then boil a potato, cut it into pieces, and distribute it to students to eat as they recited "the potato sacrament," a list of statistical mantras he hoped they would remember. They were thereby inducted into the family of class alumni.

Initially, I thought this was absurd. As a young-looking untenured professor, I worried that embracing similar antics would lead students and colleagues to write me off as unserious. I taught "seriously" for a few years, then I revived potato technology and the potato sacrament one year as a tribute to John. I used French fries instead of a boiled potato and incorporated my own mantras. I found that what I had dismissed as a gimmick was both delightful and memorable for my students. They remember the experience, and many still remember the mantras. What are we teaching for, if not for our principles to be remembered?

I now try to incorporate at least one tangible, real-world object into my teaching each week (Ho, 2026): a potato for PCA, a slinky for nonlinear transformations, a playdough mountain for maximum likelihood estimation. By far the most popular "gadget" is a handsewn normal distribution plushie (Figure 5), which I use in almost every class to

illustrate a variety of statistical and measurement concepts. In a legal deposition in 2023 (*Cayla J. v. State of California,* No. RG20084386), I held it up to illustrate the problem with proficiency (Ho, 2008), and the lawyers requested to enter it into the record as Exhibit 667. Tangible objects create memories, connections, and sometimes, unexpectedly, legal exhibits.

Figure 5. *Normal distribution plushies for making measurement memorable.*

Note. Plushies at left hand-sewn by Nicole Dick (Etsy). The top left panel illustrates the reliability coefficient as a ratio of true score variance to observed score variance. The bottom left panel shows Exhibit 667 from a legal deposition in which the author used the plushie to illustrate bias in proficiency-based gap trends (*Cayla J. v. State of California*, 2023). The right panel shows NCME-branded plushies distributed at the NCME Presidential Closing Reception, April 26, 2025.

In my measurement class, I now close the semester with an adaptation of John's potato sacrament: the Plushie Promise, where students receive their own normal

distribution plushie, then rise and recite a set of measurement mantras. At the in-person delivery of this address on April 25, 2025, I invited all attendees to rise and recite the first-ever NCME Plushie Promise (Figure 6). I later distributed hand-sewn NCME plushies at the Presidential Closing Reception the following evening. The mantras in Figure 6 encode the three claims, the two guiding principles (be precise about imprecision; be decisive amidst indecision), and the 3-4-5 triangle that this article presents. I close here as I closed my address: Commit to communication competencies. Make measurement memorable.

Figure 6. *The NCME Plushie Promise*.

On the true scores of this NCME plushie I promise that:
I will navigate the pedantry parabola with the 3-4-5 triangle.
I will ask and answer the 3 Ws.
I will not evaluate a test before identifying its quadrant.
I will remember the 5 Cs of validity evidence.
I will communicate precisely about imprecision.
With evidence, I will communicate decisively amidst indecision.
I will commit to improve my communication competencies…
and never let negative errors discourage me from raising my true score.

Note. Recited by attendees of the Presidential Address at the NCME Annual Meeting, April 25, 2025. Plushies distributed at the Presidential Closing Reception, April 26, 2025.

References

Ackerman, T., Bandalos, D., Briggs, D., Everson, H., Ho, A. D., Lottridge, S., Madison, M., Sinharay, S., Rodriguez, M., Russell, M., von Davier, A., & Wind, S. (2024). Foundational competencies in educational measurement. *Educational Measurement: Issues and Practice, 43*(3), 7-17. https://doi.org/10.1111/emip.12581

Adhvaryu, A., Kala, N., & Nyshadham, A. (2023). Returns to on-the-job soft skills training. *Journal of Political Economy, 131*(8), 2165–2208. https://doi.org/10.1086/724320

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Betebenner, D., & DePascale, C. (2024). *The time trap: Why it's misguided to report state assessment results as "years of learning."* National Center for the Improvement of Educational Assessment. https://www.nciea.org/wp-content/uploads/2024/08/TimeTrapReport-FINAL.pdf

Braun, H., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan, 86*(7), 488–497. https://doi.org/10.1177/003172170508600705

Brookhart, S. M. (2024). Commentary: Where does classroom assessment fit in educational measurement? *Educational Measurement: Issues and Practice, 43*, 18-22. https://doi.org/10.1111/emip.12626

Castillo, W., & Gillborn, S. (2025). Leveraging QuantCrit to expose and challenge systemic racism in educational psychology. *Educational Psychologist*, *60*(4), 301–316. https://doi.org/10.1080/00461520.2025.2554592

Cayla J. v. State of California, No. RG20084386 (Alameda County Superior Court 2024).

Cheng, Y., Pei, B., Filonczuk, A., & Le, A. T. (2024). Commentary: A data-driven analysis of recent job posts to evaluate the foundational competencies. *Educational Measurement: Issues and Practice, 43*, 39–44. https://doi.org/10.1111/emip.12630

Crespo Cruz, E. J., Immanuel, A., Keller, L. A., Ketan, K., McIntee, K., Serrano, F. J. M., Sireci, S. G., Smith, N., Suárez-Álvarez, J., Wells, C. S., Woodland, R., & Zenisky, A. L. (2024). Commentary: What is truly foundational? *Educational Measurement: Issues and Practice, 43*, 52–55. https://doi.org/10.1111/emip.12633

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics, 132*(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives, 36*(3), 75–102. https://doi.org/10.1257/jep.36.3.75

Fahle, E. M., Saliba, J., Kalogrides, D., Shear, B. R., Reardon, S. F., & Ho, A. D. (2024). Stanford Education Data Archive: Technical Documentation (Version 5.0). Retrieved from https://purl.stanford.edu/cs829jn7849.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests.* National Academy Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education, 21*(2), 158–179.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*(4), 709–722.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*, 351-360. https://doi.org/10.3102/0013189X08323842

Ho, A. D. (2014). Variety and drift in the functions and purposes of assessment in K-12 education. *Teachers College Record, 116*(11), 1-18. https://doi.org/10.1177%2F016146811411601116

Ho, A. D. (2016). Castles in the clouds: The irrelevance of vertical scales for most practical concerns. *Measurement: Interdisciplinary Research and Perspectives, 14,* 34-38. https://doi.org/10.1080/15366367.2016.1139983

Ho, A. D. (2022). Specifying the three Ws in educational measurement: Who uses which scores for what purpose? *Journal of Educational Measurement, 59*, 418-422. https://doi.org/10.1111/jedm.12355

Ho, A. D. (2023). Reclaiming "months of learning": An accessible metric for aggregate educational progress. Unpublished memorandum presented to the Reporting and Dissemination Committee of the National Assessment Governing Board, November 16, 2023.

https://andrewho.scholars.harvard.edu/sites/g/files/omnuum4106/files/andrewho/files/naep_months_of_learning_-_memo_to_nagb_reporting_and_dissemination_-_andrew_ho_v01.pdf

Ho, A. D. (2024). Measurement must be qualitative, then quantitative, then qualitative again. *Educational Measurement: Issues and Practice, 43*(4), 137-145. https://doi.org/10.1111/emip.12662

Ho, A. D. (2025). The conversational celebration of learning. Unpublished memo. http://bit.ly/andrewhoccl

Ho, A. D. (2026). The Five Gs for Teaching Statistics: Greek, Graphs, Grammar, Gadgets, and Games. *Teaching Educational Research Methods, 1*(1), 1-13. https://doi.org/10.5149/term.3648

Ho, A. D., & Haertel, E. H. (2007). *(Over)-interpreting mappings of state performance standards onto the NAEP scale*. Washington, DC: Council of Chief State School Officers. https://andrewho.scholars.harvard.edu/sites/g/files/omnuum4106/files/2026-05/Ho%20Haertel%20Overinterpreting%20Mappings.pdf

Ho, A. D., & Polikoff, M. S. (2025). Test-based accountability in K-12 education. In M. Pitoniak and L. Cook (Eds.), *Educational Measurement* (5th ed., pp. 1113-1188). Oxford University Press. https://doi.org/10.1093/oso/9780197654965.003.0016

Ho, A. D., Reardon, S. F., & Kalogrides, D. (2021). Validation methods for aggregate-level test scale linking: A rejoinder. *Journal of Educational and Behavioral Statistics, 46*, 209-218. https://doi.org/10.3102%2F1076998621994540

Ho, A. D., Ackerman, T., Bandalos, D., Briggs, D., Everson, H., Lottridge, S., Madison, M., Sinharay, S., Rodriguez, M., Russell, M., von Davier, A., & Wind, S. (2024). Foundational competencies in educational measurement: A rejoinder. *Educational Measurement: Issues and Practice, 43*(3), 56-63. https://doi.org/10.1111/emip.12623

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GEs. *Educational Measurement: Issues and Practice, 3*, 8–14.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 221–233.

Jacobs, W. W. (1906). The monkey's paw. In The lady of the barge (6th ed.). Harper & Brothers. (Original work published 1902)

Lee, S. J. (2009). *Unraveling the "model minority" stereotype: Listening to Asian American youth* (2nd ed.). Teachers College Press.

MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics, 232*, 272–299.

Middlebrook, K., Hamilton, L. S., & Walker, M. E. (2024). Commentary: How research and testing companies can support early-career measurement professionals. *Educational Measurement: Issues and Practice, 43*, 45–48. https://doi.org/10.1111/emip.12631

Milner, H. R. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher, 36*(7), 388–400. https://doi.org/10.3102/0013189X07309471

Moses T., & Dorans, N. J. (2021). Aggregate-level test-scale linking: A new solution for an old problem? *Journal of Educational and Behavioral Statistics*, 46(2), 187–202.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton University Press.

Quinn, D. M., & Desruisseaux, T. M. (2022). Replicating and extending effects of "achievement gap" discourse. *Educational Researcher*, *51*(7), 496-499. https://doi.org/10.3102/0013189X221118054

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics, 46*(2), 138–167. https://doi.org/10.3102%2F1076998619874089

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & Saliba, J. (2024). Stanford Education Data Archive (Version 5.0). Retrieved from https://purl.stanford.edu/cs829jn7849.

Shultz, P. K., & Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. *Frontiers in Education, 6*, Article 701973. https://doi.org/10.3389/feduc.2021.701973

Spencer Foundation Task Force on Preparation for Transformative Research. (2025). *Enhancing the preparation of researchers for transformative research in education*.

The Spencer Foundation. https://spencerfoundation.s3.us-east-2.amazonaws.com/store/148451459b3c79c804642fb802d6589c.pdf

van Orman, D. S. J., Jackson, J. A., Vo, T. T., & Taylor, D. D. (2024). Commentary: Perspectives of early career professionals on enhancing cultural responsiveness in educational measurement. *Educational Measurement: Issues and Practice, 43*, 27–32. https://doi.org/10.1111/emip.12628

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press.